一种基于MDL度量的选择性扩展贝叶斯分类器

王 峻[1],周孟然[2]

(1.淮南师范学院,安徽 淮南 232001;2.安徽理工大学,安徽 淮南 232001)

摘 要:朴素贝叶斯分类器是一种简单而高效的分类器,但它的条件独立性假设使其无法表示属性间的依赖关系。TAN分类器按照一定的结构限制,通过添加扩展弧的方式扩展朴素贝叶斯分类器的结构。在TAN分类器中,类变量是每一个属性变量的父结点,但有些属性的存在降低了它分类的正确率。文中提出一种基 于MDL度量的选择性扩展贝叶斯分类器(SANC),通过MDL度量,删除影响分类性能的属性变量和扩展弧。实验结果表明,与NBC和TANC相比,SANC具有较高的分类正确率。

关键词:朴素贝叶斯;贝叶斯网络;MDL;度量

中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2007)07-0035-03

A Selective Augmented Naive Bayesian Classifier Based on MDL Score

WANG Jun[1], ZHOU Meng-ran[2]

(1.Huainan Normal University, Huainan 232001, China;
2.Anhui University of Science and Technology, Huainan 232001, China)

Abstract: Naive Bayesian classifier is a simple and effective classifier, but its conditional independence assumption makes it unable to express the dependence among features. TAN classifier extends the structure of Naive Bayes classifier by adding augmenting arcs that obey certain structural restrictions. In TAN classifier, all features are constrained to have the class variable as a parent, but some features degrade its classification accuracy. The present paper presents SANC(A Selective Augmented Naive Bayesian Classifier based on MDL score) that removes features and augmenting arcs which affect the performance of classification by MDL score. Compared with NBC and TANC, experimental results show SANC has higher accuracy.

Key words: naive Bayes; Bayesian network; MDL; score

0 引 言

分类是机器学习和模式识别的基本问题,提高分类器的性能是分类研究的主要目标之一。在众多的研究建立分类器的方法和理论中,建立在直观的概率论基础上的朴素贝叶斯方法(Naive Bayes,NB)以能够将先验信息和样本信息进行集成等优点,广泛应用于许多领域,其性能可以与神经网络、决策树相媲美[1,2]。但朴素贝叶斯分类器是基于条件独立性假设,在实际应用中,各属性变量之间常常具有明显的依赖性,因此可以通过借鉴贝叶斯网络中表示属性依赖关系的方法,通过在相关属性间添加扩展弧的方式,扩展朴素贝叶斯的结构,使其能容纳属性间存在依赖关系。Friedman[1,3]研究了具有树结构的TAN(tree augmented naive Bayes)分类器,BAN(Bayesian network augmented naive Bayes)[1]进一步扩展了TAN的结构,允许属性之间可以形成任意的有向图,使其表示属性间依赖关系的能力增强。

笔者在研究TAN分类器的基础上,提出一种基于MDL度量的选择性扩展贝叶斯分类器,并通过实验与NBC、TANC进行比较,分析该方法在分类精度上的效果。

1 树扩展朴素贝叶斯分类器TAN

贝叶斯网络是一个带有概率注释的有向无环图,由网络的拓朴结构 G 和局部概率分布的集合 θ 两部分组成,可表示为 $B=(G,\theta)$,G 中结点表示知识领域的随机变量,有向弧表示变量间的因果关系,θ 是量化网络的一组参数,表达各结点间因果关系影响的强度。

收稿日期:2006-11-16

基金项目:安徽省高等学校省级自然科学研究项目(KJ2007B075)

作者简介:王 峻(1967-),男,安徽淮南人,硕士,讲师,研究方向为数据挖掘;周孟然,博士,教授,研究方向为计算机控制。

已经证明：完全的贝叶斯网络结构的学习是 NP 问题[4]。而 TAN 是一种受限制的贝叶斯网络模型，是严格限制的朴素贝叶斯分类器与无限制的贝叶斯网络之间的一种折衷。

令 $D = \{A_1, A_2, \cdots, A_n, C\}$，其中 $A_1, A_2, \cdots, A_n$ 是属性变量，a_i 是属性 A_i 的取值，$x_i = (a_1, a_2, \cdots, a_n)$ 是各个实例；$\prod A_i (i = 1, 2, \cdots, n)$ 是 A_i 的父结点集，$|\prod A_i| \leqslant 2$；C 是类变量，$\{c_1, c_2, \cdots, c_l\}$ 是类变量 C 的可能取值。根据贝叶斯最大后验准则，TANC 的输出类别标号为：

$$C_{\text{TANC}} = \arg\max_{cj \in C} p(cj \mid a_1, a_2, \cdots, a_n) =$$

$$\arg\max_{cj \in C} \frac{p(a_1, a_2, \cdots, a_n \mid cj) p(cj)}{p(a_1, a_2, \cdots, a_n)} =$$

$$\arg\max_{cj \in C} p(a_1, a_2, \cdots, a_n \mid cj) p(cj) =$$

$$\arg\max_{cj \in C} p(cj) \prod_{i=1}^{n} p(a_i \mid \prod A_i)$$

TANC 构造算法如下[4,5]：

(1) 通过训练集计算属性对之间的相关性函数 $f(A_i, A_j, D)$ $(i \neq j)$。

(2) 建立一个以 $A_1, A_2, \cdots, A_n$ 为结点的加权完全无向图，结点 A_i, A_j 之间的权重为 $f(A_i, A_j, D)$ $(i \neq j)$。

(3) 利用求最大权生成树算法，建立该无向图最大权重跨度树。首先把边按权重由大到小排序，之后遵照被选择的边不能构成回路的原则，按照边的权重由大到小的顺序选择边，这样由所选择的边构成的树便是最大权重跨度树。

(4) 指定一个属性结点作为根结点，将所有边的方向设置成由根结点指向外，把无向图转换成有向图。

(5) 加入类结点 C，并添加从 C 指向每个属性结点 A_j 的弧。

该算法的时间复杂度为 $O(n^2 N)$，其中 n 是属性的个数，N 是训练实例的个数。

2 基于 MDL 度量的 SANC

贝叶斯网络结构学习算法分为两类：一类是基于相关性分析的方法，用互信息度量表达结点之间的相关性；另一类是基于打分和搜索的方法，常见的打分函数有 Bde 度量、BIC 度量和 MDL 度量，文中提出的选择性扩展贝叶斯分类器是基于 MDL 度量。

2.1 MDL 度量的原理

最小描述长度(MDL)原理是 Rissanen[5]在研究通用编码时提出的。其基本原理是对于一组给定的实例数据 D，如果要对其进行保存，为了节省存储空间，一般采用某种模型对其进行编码压缩，然后再保存压缩后的数据。同时，为了以后正确恢复这些实例数据，将所用的模型也保存起来。所以需要保存的数据长度(比特数)等于这些实例数据进行编码压缩后的长度加上保存模型所需的数据长度，将该数据长度称为总描述长度。最小描述长度(MDL)原理就是要求选择总描述长度最小的模型。

如果将贝叶斯网络作为对实例数据进行压缩编码的模型，MDL 原理就可以用于贝叶斯网络的学习。该度量被视为网络结构的描述长度和在给定结构下样本数据集的描述长度之和。一方面，用于描述网络结构的编码位随模型复杂度的增加而增加；另一方面，对数据集描述的编码位随模型复杂度的增加而下降。因此，贝叶斯网络的 MDL 总是力求在模型精度和模型复杂度之间找到平衡。构建贝叶斯网络首先定义一个评分函数，该评分函数描述了每个可能结构对观察到的数据拟合，其目的就是发现评分最大的结构，这个过程连续进行到新模型的评分分数不再比老模型的高为止。

对于贝叶斯网络 $B = (G, \theta)$，贝叶斯网络结构的 MDL 度量由描述贝叶斯网的数据长度和实例数据的压缩长度两部分构成[6]。贝叶斯网的描述长度为：无向图的编码长度为 $\sum_i (1 + |\prod A_i|) \log n$ 比特，条件概率表的数据长度为 $\frac{1}{2} \log N \sum_i |\prod A_i| (|A_i| - 1)$，所以贝叶斯网络结构的描述长度为

$$\text{MDL}_{\text{net}} = \sum (1 + |\prod A_i|) \log n + \frac{\log N}{2} \sum_i |\prod A_i| (|A_i| - 1)$$

数据的压缩长度为：$\text{MDL}_{\text{data}} = N \sum_i H(A_i \mid \prod A_i)$，其中，$H(X \mid Y) = -\sum P(X, Y) \log P(X \mid Y)$ 为变量 X 对变量 Y 的条件熵，而 X 与 Y 的互信息度量 $I(X, Y) = \sum P(X, Y) \log \frac{P(X, Y)}{P(X) P(Y)}$，由 $H(X \mid Y) = H(X) - I(X, Y)$ 可得 $\text{MDL}_{\text{data}} = N \sum_i H(A_i) - N \sum_i I(A_i, \prod A_i)$，由于 $N \sum_i H(A_i)$ 在描述所有结构长度时是常量，可以忽略，因此贝叶斯网络结构的整体描述长度为：

$$\text{MDL}(B \mid D) = \text{MDL}_{\text{net}} + \text{MDL}_{\text{data}} = \sum_i (1 + |\prod A_i|) \log n + \frac{\log N}{2} \sum_i |\prod A_i| (|A_i| - 1) - N \sum_i I(A_i, \prod A_i) \quad (1)$$

2.2 SANC

构建选择性扩展贝叶斯分类器的关键之一是合理地选择属性变量。在 TAN 分类器中，将所有的属性变量均作为类变量的孩子结点。而在选择性扩展贝叶斯

分类器中,并不是所有的属性变量都作为类变量的孩子结点,而是只将相关属性变量作为类变量的孩子结点,对于那些可能降低分类器的分类性能的不相关的属性,将其从分类器结构中删除;构建选择性贝叶斯分类器的关键之二是合理地选择扩展弧,通过合理地选择扩展弧求出最佳的分类器结构。

对于在选择性扩展贝叶斯分类器中添加的任意有向弧 (A_i, A_j),可通过下式计算评分函数的增量:

$$W(A_i, A_j) = \text{score}(A_j, \prod A_j = \{A_i, C\}) - \text{score}(A_j, \prod A_j = \{C\})$$

由式(1)得

$$W(A_i, A_j) = \log n + \frac{\| C \| (\| A_i \| - 1)(\| A_j \| - 1)\log N}{2} - NI(A_j, A_i \mid C)$$

$$W(C, A_j) = \log n + \frac{(\| C \| - 1)(\| A_j \| - 1)\log N}{2} - NI(A_j, C)$$

选择性扩展贝叶斯分类器结构算法如下:

1) 合理选择属性变量。对于属性集中任意属性变量 A_j,如果满足 $W(C, A_j) = \text{score}(A_j, \prod A_j = \{C\}) - \text{score}(Y, \prod A_j = \emptyset) < 0$ and $\forall A_i, W(A_i, A_j) < 0$,将 A_j 删除,依次计算求出合理的属性子集;

2) 在属性子集中的各个属性变量之间建立完全有向图;

3) 用 $W(A_i, A_j)$ 评价每一个扩展弧,若 $W(A_i, A_j) < 0$,将该扩展弧删除;

4) 加入类结点 C,并添加从 C 指向属性子集中的每个属性结点 A_j 的弧。

3　实验结果及分析

3.1　实验结果

本实验所用的数据来自 UCI 机器学习数据库,从中选择四个数据集分别为:Vote, Tic－Tac－Toe, Zoo, Postoperative－Patient。表 1 列出了每个数据集的实例个数、类个数、属性个数。所有实验是在 Weka[7] 系统中完成的,在测试过程中采用 10 次交叉验证法估计分类器的正确率。实验结果如表 1 所示。

3.2　实验结果分析

实验的主要目的是对 NBC, TANC 和 SANC 分类器在每个数据集上的分类正确率进行比较,每个分类正确率是在测试集上成功预测的实例数占总实例数的百分比。实验结果表明,尽管在 Tic－Tac－Toe 数据集上,SANC 的分类正确率有所下降,但在其它 3 个实验数据集上均取得了较好的分类性能,其分类正确率优于 NBC 和 TANC,从而表明了 SANC 分类器的有效性。

表 1　两种分类器的实验结果

数据集	实例数	属性个数	分类正确率(%)		
			NBC	TANC	SANC
Vote	435	16	75.2	77.6	86.7742
Tic－Tac－Toe	958	9	69.3835	77.325	72.6228
Zoo	101	17	93.0693	97.0297	99.0099
Postoperative－patient	90	8	67.7778	67.7778	68.8889

4　结束语

提出一种选择性扩展贝叶斯分类器,通过 MDL 度量选择合理的属性子集和扩展弧,实验表明,该分类器具有较好的分类正确率。但是由于 MDL 没有用到先验知识,其学习结果的正确性完全依赖于样本数据集合,因此是否有更好的度量方法应用于选择性扩展贝叶斯分类器以及针对不同的数据集,能否有选择地选择分类器以达到最佳的分类效果,这些都是下一步研究工作的重点。

参考文献:

[1] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifier[J]. Machine Learning, 1997(29):131－163.

[2] Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers[C]// In: Rosenbloom P, Szolovits P. Proc. of the 10th National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 1992:223－228.

[3] Friedman N, Goldszmidt M. Building classifiers using Bayes network[C]//In Proc . National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1996:1227－1284.

[4] Chickering D, Geiger D, Heckerman D. Learning bayesian networks in np－hard[R]. MSR－TR－94－17. US: Microsoft, 1997.

[5] Stochastic R J. Complexity in Statistical Inquiry[M]. Singapore:[s. n.], 1989.

[6] Friedman N, Goldszmidt M. Discretization of continuous attributes while learning bayesian networks[C]// In Proceedings of the Thirteenth International Conference on Machine Learning. [s. l.]:[s. n.], 1996:157－165.

[7] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations[M]. Seattle: Morgan Kaufmann Publishers, 2000:265－314.